METROPOLITAN HOUSING AND COMMUNITIES POLICY CENTER

# Performance Measurement to Evaluation

Sponsored by the World Bank Group Community Outreach Program

*Peter A. Tatian*

*March 2016*

**Nonprofits are increasingly expected to use data and evidence to manage and improve their programs and to demonstrate that they are producing positive results for the clients they serve. A growing number of national foundations are incorporating an array of approaches to assessing grantmaking performance: dashboards, scorecards, results frameworks, and more formal evaluations (Putnam 2004, 17–20). At the federal level, the Government Performance and Results Modernization Act of 2010 (GPRMA) requires departments and agencies to establish annual performance assessments of government programs (Dodaro 2011). Nonprofits that receive federal grant funding may have to report performance data to meet GPRMA requirements.**

Two frameworks for using data and evidence are performance measurement and evaluation. Though some may use these terms interchangeably, they represent distinct approaches that have different objectives. In short, the main distinctions are as follows:

- Performance measurement tells what a program did and how well it did it.
- Evaluation tells the program's effect on the people, families, or communities it is serving, that is, whether a program is producing results or having an impact (box 2).

BOX 1

### Measure4Change

Measure4Change is a program of the World Bank Group and the Urban Institute to build performance measurement capacity among local nonprofits in the Washington, DC, metropolitan area. Nonprofits recognize the importance of measuring program effectiveness, but their abilities vary, and resources for improvement are scarce. Measure4Change aims to fill this long-standing gap between what nonprofits in the DC metropolitan area want and what they are able to do. The effort intends to deliver performance measurement training in a way that is practical and accessible for nonprofits and over an extended period of time to help it take hold. The ultimate goal of this effort is to help the DC region's nonprofits better understand how they are helping their constituencies and how they can do better. Measure4Change, sponsored by the World Bank Group, has three components: grant support and one-on-one technical assistance for grantees, a regional community of practice, and knowledge briefs.

But, how should nonprofits use performance measurement and evaluation to inform their work? What specific methods should they use and when? Measure4Change, a collaboration between the World Bank Group and the Urban Institute, seeks to increase the performance measurement and evaluation capacity of nonprofits in the Washington, DC, area and inform nonprofits and funders vexed by such questions. This brief sets forth recommendations to give nonprofits, funders, and others a basic overview of performance measurement and evaluation and help them assess which activities are beneficial for programs at particular stages of development. The brief introduces the idea of a *performance measurement-evaluation continuum*, which lays out a progression first using performance measurement and formative evaluation to improve and refine programs, then, for selected programs if and when appropriate, undertaking summative evaluation to determine impact.

This brief first discusses the performance measurement side of the continuum, then evaluation. The final section presents the continuum itself and describes how the different components ideally align and build off each other. The framing for this brief is performance measurement and evaluation of an *individual program*—a defined set of activities or interventions intended to produce specific outcomes for a particular population. Though this brief frames the discussion in terms of a single program for simplicity, these same concepts may be adapted to larger initiatives or to entire organizations.

Another challenge for practitioners is that the use and definition of performance management and evaluation terms can vary widely, even among experts in the field. This can make it difficult to apply these concepts in practice, as not all parties may be "speaking the same language." To help alleviate miscommunication, the brief includes a glossary of performance measurement and evaluation terminology to clarify how this brief uses specific terms. Where possible, these definitions are taken from published resources, but readers should be aware that not all sources will agree on the specific meanings used here.

BOX 2

**What Is Impact?**

The term *impact* is sometimes used rather loosely to describe any change observed in program participants, but, most properly, *impact* should refer to the net effect of a program relative to what would have happened had the program not existed. In other words, the impact should be changes in outcomes *produced by the program alone* and not caused by other factors. Impact most typically reflects direct outcomes of a program, that is, persistent changes in participants' situations or behavior. But, impacts can also include changes outside of those seen in program participants, such as impacts on other family members or the larger community. Impacts can also be either intended (i.e., those meant to be caused by the program) or unintended (i.e., those incidental to the program's original goals or objectives).

# Performance Measurement

Performance measurement involves collecting and reporting data that can be used to summarize and assess the way a program is being implemented. Performance measurement is intended to generate information a nonprofit can use to improve a program, a process often described by the phrase *continuous improvement*. Typically, performance measurement data are collected with some frequency and immediacy. That is, performance measurement data, to best improve program performance, are collected and assessed while participants are being served, rather than only after a program cycle has concluded. Data are generally collected in the same way for future participants so that progress can be tracked over time.

Implicit in performance measurement is the idea of performance management, in which data are actively used to revise an ongoing program to improve efficiency or results. Performance management is a "dynamic process" intended to "produce better outcomes for participants."[1] Typically, performance management will include processes that supply data in actionable forms (such as dashboards or easily digestible reports) to key personnel—from organization boards to senior management to frontline staff—as well as opportunities to collectively analyze and make decisions based on these data.

To be most effective, performance measurement should be based on an underlying *theory of change*, a program's approach to altering something about the world. A theory of change can be understood as a conceptual road map that outlines how a series of actions can bring about a desired outcome. For example, an out-of-school program that uses sports to improve academic outcomes might say that developing teamwork skills will lead to better self-discipline and confidence that will, in turn, improve young children's classroom performance.[2]

A theory of change is often articulated more formally in the form of a *logic model*, a graphic that shows how a program is intended to work and achieve results. A logic model depicts a path toward a goal. It will make clear what specific resources need to be in place, what activities need to happen, and what changes  (in both the near and long terms) will ultimately lead to desired outcomes. The logic model will also specify what measurements should be taken (i.e., what data should be collected) to confirm that appropriate and expected progress is being made at different points along this path.[3]

These data points become the basis for performance measurement and management. Box 3 shows types of basic questions that performance measurement data can answer. For example, the program's theory of change/logic model should indicate what particular population would benefit. The nonprofit can therefore collect data on participant characteristics to confirm that the program is indeed serving this desired population. Similarly, the theory of change/logic model may say that a certain level of program exposure (such as a number of hours of participation) is necessary to produce an outcome. The nonprofit should also collect data on the services participants receive to ensure each is receiving the needed amount of classes, hours, or the like.

BOX 3

**What Questions Can Performance Measurement Data Answer?**

- Inputs
  - » What staff/volunteers are involved in the program?
  - » What is the program budget?
  - » What equipment and materials does the program have?
  - » How were all the inputs used?
- Program participation
  - » Who is participating (i.e., participant characteristics)?
  - » How many participants are there?
  - » How much did each person participate (especially relative to a desired level of service)?
- Outputs
  - » What services are delivered?
  - » Who delivered the services?
  - » How well are services being delivered?
- Outcomes
  - » What changes do we observe in participants (generally restricted to changes under the direct control of or directly affected by the program)?

One approach used in performance measurement is the *pre-post analysis*, in which outcomes are measured for participants at the beginning and at the end of their involvement in a program.[4] Differences between the "pre-program" and "post-program" results can be considered partial evidence that a program is helping participants achieve better results. As is discussed in the next section on evaluation, however, a pre-post analysis, by itself, cannot determine a program's impact.

Armed with performance measurement information, nonprofits can examine performance relative to a program's theory of change/logic model. Is the program working the way the nonprofit would expect or want? If not, where are things breaking down? What needs to be changed so that better results can be achieved? If, for instance, the right people (as defined in the theory of change) are not being served by a program, perhaps better outreach would attract more appropriate participants. If too many participants drop out of a program prematurely, more effective retention strategies may be required. These data and questions need to be reexamined repeatedly so that ongoing refinements (i.e., continuous improvement) can be made.[5]

# Evaluation

*When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative.*

*—Robert Stakes, as quoted in Allen Nan, "Formative Evaluation," Beyond Intractability, December 2003, http://www.beyondintractability.org/essay/formative-evaluation*

Through effective performance measurement, program staff can collect much valuable data—information that can help improve service delivery and client results. Performance measurement data cannot, however, directly answer questions about program impact—whether the program alone produced any observed outcomes. Nor can performance measurement data necessarily answer all questions about how a program is working or how results were achieved. To answer these questions, evaluation methods are needed. (For more on the differences between performance measurement and evaluation, see box 4.)

Evaluation covers a variety of activities that provide evidence about what a program did and whether (or how well) it achieved its aims (box 5). While these evaluation methods, if done well, can provide valuable information about a program, not all measure impact directly or provide generalizable conclusions about whether a program was effective. It is important to understand what questions a particular type of evaluation can answer. Further, evaluation methods should be viewed as complementary—using multiple methods can give a more complete picture of a program.

BOX 4

### What Is the Difference between Performance Measurement and Evaluation?

The dividing line between performance measurement and evaluation, particularly formative evaluation, can be blurry, with some activities and goals common to both. But, a few things generally distinguish evaluation from performance measurement.

*Performance measurement is ongoing, while evaluation is discrete*. Performance measurement is part of a continuous improvement process, in which data are collected, analyzed, and reported as close to real time as possible, giving staff immediate and actionable feedback on a program's status. Evaluation, whether formative or summative, is not done continuously but rather during particular periods of a program's development or implementation and over a specified timeframe. For example, a formative evaluation may be carried out during the first six months of a program's planning and implementation, while a summative evaluation might be done during the fifth year of an established program.

*Performance measurement is responsive and adaptive; evaluation answers a predetermined set of questions.* While performance measurement is intended to answer questions about a program's execution, these questions are not fixed and can change as the program evolves. Performance measurement data themselves may suggest additional questions to be answered and new data to be collected. In contrast, an evaluation starts with a set of questions and then uses appropriate methods to answer those specific questions. While questions may be adjusted somewhat (or added to) during the evaluation, generally they are not changed substantially.

*While performance measurement exploits program and outcome data, which can also be used in evaluation, evaluation usually involves other data collection and research methods.* Performance measurement uses data that can be collected routinely during program operations, such as clients' use of services, and assessment tools that measure outcomes. An evaluation will often expand on those data sources, however, by collecting additional data through surveys, direct observation, or other means. Furthermore, evaluations frequently use qualitative methods, such as interviews and focus groups, to gather information about client or staff experiences.

*Performance measurement is mostly done by program staff, whereas evaluation is typically carried out by people outside the program.* While this is not a hard and fast rule, program staff are most closely engaged in collecting and examining performance measurement data, although larger nonprofits may have performance measurement specialists or a team that assists with this work. In contrast, program evaluations are often conducted by an outside evaluator. The evaluator might be a completely separate individual or organization hired by the nonprofit, or it may be a distinct evaluation unit within the nonprofit itself. The main reasons for using an outside evaluator are to get a more objective assessment than might be provided by internal program staff or to access skills and expertise not available within the nonprofit.[a]

In addition, as discussed in box 6, certain types of evaluation are capable of determining program impact, which performance measurement cannot do directly.

[a] For more on working with external evaluators, see Harlem Children's Zone (2012).

BOX 5

**What Are the Types of Evaluation?**

- Formative evaluation
  - » Planning study
  - » Process study
- Summative evaluation
  - » Experimental study
  - » Comparison study

Evaluation results can be used to help nonprofits decide whether programs might be successfully scaled up to serve more people similar to those already being served. The ability to draw this conclusion depends on an evaluation's *internal validity*, that is, how accurately an evaluation measures a program's impact on outcomes for the population it is serving. In addition, evaluation results may help nonprofits decide whether to expand the program to other locations or to different populations. The strength of the results in answering this question depends upon the evaluation's *external validity*, that is, the extent to which conclusions about program impact can be reliably generalized to other populations, programs, geographies, or time spans than those studied. Will a tutoring program that was successful for teenage boys work equally well for teenage girls or younger boys? Can a successful job training program in Cleveland be replicated in Denver? The program's theory of change/logic model can help nonprofits assess the external validity of a program's evaluation results for other populations or in different environments.

More detail on the different types of evaluations is provided below, including how well they answer particular questions about programs. Not all evaluation methods are suitable for all programs. The next section will discuss this further in the context of the performance measurement-evaluation continuum.

## Formative Evaluation

The purpose of *formative evaluation* is to learn how a program is being designed or carried out, with the objective of providing information that can be used to improve implementation and results.[6] One type of formative evaluation, a *planning study*, takes place during the design or planning phase to clarify a program's plans and to make improvements at an early stage. Questions addressed by a planning study can include the following:

- What are the goals and objectives of the program?
- What population is the program intended to serve?
- Is the program intervention appropriate for the identified goals and population?

- What impact is the program expected to have? Is there sufficient evidence to support this predicted impact?
- Are the available resources (staff, facilities, equipment, funding) adequate to accomplish the program's goals and objectives?
- Is the program's implementation timeline achievable?

In addition, a planning study can lay the groundwork for future formative and summative evaluations by developing program indicators and benchmarks.[7]

Formative evaluation can also be undertaken throughout program implementation, as an *implementation* or *process study*. A process study can be particularly important for programs that are still developing, so that changes during the implementation phase can be clearly documented. A process study[8] answers questions such as the following about the quality of program implementation:

- What interventions were implemented?
- Did services get delivered as intended? Were the appropriate populations reached by the program? If not, why not?
- Were the resources (staff, facilities, equipment, funding) sufficient to accomplish the program's goals and objectives?
- Did staff encounter problems in setting up or running the program? Were they able to respond to and address all challenges?

A process study can help a nonprofit understand why a program is getting—or not getting—particular results and provide information leadership and staff can use to change the program as it is being implemented. Data collected during process studies can also be invaluable for subsequent summative evaluations that assess program impact.[9]

Formative evaluations typically use a variety of research methods to achieve their objectives. Certainly, performance management data can be a valuable input to process studies. Evaluators may also want to collect additional data through surveys, direct observation of clients or staff, or other means, to supplement routine performance management data. Organizations will most likely analyze these data using descriptive methods, such as creating summary tables or charts, rather than the multivariate methods common in summative evaluations. Further, formative evaluations often employ qualitative research methods, such as interviews or focus groups with clients or staff, to gather more descriptive data on how well a program is serving clients and meeting its goals and objectives.

## Summative Evaluation

While formative evaluations examine current program operations and results, *summative evaluations* look retrospectively at what a program accomplished. The purpose of a summative evaluation is to learn how effectively a program has changed the conditions described in the theory of change/logic model—

specifically, to determine the program's impact on those conditions (see box 6). A summative evaluation can help funders determine whether additional resources to serve more people would be a good investment.[10] On the other hand, nonprofits or funders might decide that a program failing to show positive impact should be redesigned or phased out.

This section discusses two types of summative evaluations intended to measure impact: experimental studies and comparison studies.

BOX 6

### How Do Evaluators Measure Impact?

Why can't evaluators just use performance measurement or similar data to assess impact? The answer concerns a fundamental component of evaluation, the *counterfactual*. A counterfactual is a set of conditions that characterize how participants would have fared had a program not existed. The counterfactual is crucial because changes observed in participant outcomes cannot be entirely attributed, on their own, to the effects of the program. For one thing, some participants may have improved their outcomes even without the program's help. For a job training program, for instance, some participants may have found jobs on their own regardless of whether they received assistance, so the program cannot legitimately take full credit for those successes.

Having a clear counterfactual is especially crucial for programs trying to overcome adverse conditions. For example, during the housing crisis, many homeowners were losing their homes through foreclosure. Housing counseling programs worked with homeowners to help them find ways to stay in avoid foreclosure, if possible. But, circumstances beyond the control of the housing counselor or the homeowner, such as difficult economic conditions, may have made some foreclosures impossible to avoid. If evaluators looked only at the foreclosure rate for counseled homeowners, they might conclude that counseling was not helping. By comparing foreclosure rates to a similarly situated group of homeowners who did not receive counseling, however, evaluators might be able to show that foreclosure rates for counseled homeowners were lower than those of noncounseled homeowners, which would be a positive impact.[a]

[a] Mayer and his coauthors (2012) describe a comparison study that evaluated outcomes for homeowners who received counseling through the National Foreclosure Mitigation Counseling Program.

An *experimental study* (also referred to as a *randomized controlled trial*) is generally considered the most reliable way to determine a program's impact. In an experimental study, program participants (the treatment group) and nonparticipants (the control group) are selected at random from the same population. This eliminates the risk of *self-selection bias*—the possibility that people who choose to participate will differ in some way from those who do not—since program participation is not, in the end, left up to individuals or even the program staff to decide.

Properly done experimental studies have a high degree of *internal validity* in measuring program impacts; that is, they can accurately determine impact for populations the program is serving. Does this mean that organizations should only do experimental studies? Despite their virtues, experimental studies may not be possible or desirable for particular programs. If, for example, the program is serving a small population, sufficient numbers of treatment and control group members may not be available. Additionally, an experimental study may not be desirable for ethical reasons. To obtain a proper control group, people who need and are eligible for program services must be denied those services for the duration of the study. For some nonprofits or certain services, denying assistance is not acceptable.

For these reasons, experimental studies are most commonly applied to social service programs when programs are *oversubscribed*, that is, more people need and want services than there are program resources to help. Since people will have to be denied services anyway, there may be fewer ethical objections to randomly refusing those services as part of an experimental study.[11] Indeed, random selection may be the fairest method to decide who should benefit from a program with limited capacity. (Although people in the control group, once denied services as part of the evaluation, must remain ineligible to reapply for and receive program services until the study is completed.)

Where random selection is not possible, a *comparison study* (also referred to as a *quasi-experimental study*) may be a viable alternative. In a comparison study, changes in outcomes for program participants (the treatment group) will be compared to changes for a comparison group that resembles the program participant group as much as possible. For example, a program to help public housing residents increase their savings might look at savings rates for program participants versus other public housing residents who choose not to take part. Generally, statistical methods (such as regression models) would be used to adjust for any differences between the two groups that might affect the outcome of interest, such as household income, employment status, and educational attainment.

How reliably a comparison study estimates program impacts depends upon how well differences between the treatment and comparison groups are minimized, either through the two groups' composition or through statistical methods. Differences include not only *observable differences*, ones evaluators can measure, but also *unobservable differences* that cannot be measured and that might also affect outcomes. If, for instance, participants in a public housing savings program start out being more motivated to save money than residents who did not sign up, then this unobserved difference could well explain any differences in savings rates between participants and nonparticipants. It might be possible to measure a person's motivation to save money by asking program participants and comparison group members a series of questions. But, it will not generally be possible to measure all characteristics that might affect the outcomes being studied. This, again, highlights one of the key advantages of experimental studies—they minimize differences in both observable and unobservable characteristics between treatment and control groups.

In both experimental and comparison studies, a valid summative evaluation can help determine a program's effect on the people, families, or communities it serves and its results or impact. In other words, a summative evaluation answers a fairly straightforward question: did the program work and if so, how well? Depending on how the summative evaluation is structured, however, the results will not necessarily explain how or why a program worked (or did not work) in a way that would inform any improvements. Performance measurement data and formative evaluation findings may be better able to help answer those questions.

With a successful summative evaluation that has demonstrated positive program impact, a nonprofit may be able to do a *cost-benefit analysis* to determine the program's return on investment. A cost-benefit analysis would determine the total costs—including labor, facilities, equipment, and other direct and indirect costs—of providing program services to an individual. Then, the analysis would compare total costs to the monetized value of the changed outcomes that are program impacts. For example, a cost-benefit analysis of a job training program might compare the costs of providing services with the eventual savings in unemployment insurance based on the program's impact: increasing participants' rates of employment.

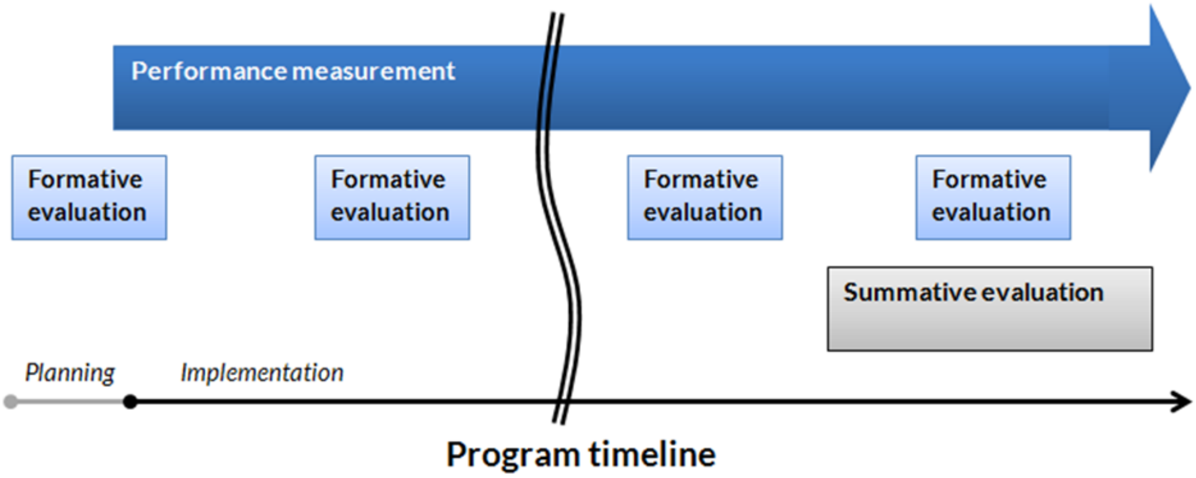
Nevertheless, an experimental or comparison study only determines impact for clients who were served in a particular community and context. These findings may not necessarily have much external validity for other clients or circumstances. Other formative evaluation data may be able to establish whether the same impacts might be realized for a program model helping different populations or working in other locations.

## Performance Measurement-Evaluation Continuum

Both performance measurement and evaluation have value in improving and determining the effectiveness of programs. Nonprofits should apply the appropriate methods to their programs based on their needs, the questions they want to answer, and the programs' stages of development. Ideally, multiple methods will be used in a complementary fashion to provide a more complete understanding of the program. To help nonprofits, and their funders, think about the roles of different performance measurement and evaluation approaches a *performance measurement-evaluation continuum* is illustrated in figure 1.

FIGURE 1

**Performance Measurement-Evaluation Continuum**



The continuum encompasses all approaches discussed in this brief and lays out an ideal sequence of performance measurement and evaluation activities. At the early planning stages, a formative (planning) evaluation would answer basic questions about the program's design and proposed implementation. This information can be used to make early adjustments for implementation.

Once the program begins operating, staff can start collecting performance data to track actual implementation and service use. As the continuum shows, performance measurement activities continue throughout program implementation as part of the continuous improvement process.

Periodically, programs would undergo additional formative (process) evaluations to answer specific questions about the quality of the program and the results being achieved. Each subsequent formative evaluation can provide management and program staff with fresh insights on how well program objectives and goals, as articulated in the theory of change/logic model, are being attained. This information can also contribute to the program's continuous improvement process.

This pattern of ongoing performance measurement and periodic formative evaluation can continue indefinitely. The frequency of formative evaluation cycles can depend on several factors. Programs in early stages of implementation, particularly those with unproven designs as well as proven programs being introduced to new populations or communities, may initially require more frequent assessment, perhaps as often as quarterly. Being so frequent, these early evaluations may not be extremely comprehensive, and each subsequent evaluation cycle may focus on different aspects. As programs mature and become more stable, such frequent evaluation may not be necessary. At this stage, annual or biennial formative evaluations, perhaps in greater depth than early evaluations, may be sufficient.

Later in the continuum, a program may reach the stage where a summative evaluation would be appropriate. A program may take several years to arrive at this point and, in reality, many programs may never reach a situation in which a summative evaluation is feasible or useful. How can nonprofits, and funders, determine whether a program is ready for a summative evaluation? A number of "evaluability" standards exist that provide guidance on this question. Several criteria can be used to assess a program's summative evaluation readiness:[12]

- *Relevance and importance of program impact question.* As noted throughout this brief, a summative evaluation fundamentally addresses the question of program impact. Before proceeding, the nonprofit and its funders should establish that answering this question is relevant to the program and likely to be of high value to the community being served.

- *Applicability, scalability, and replicability.* Because summative evaluations have high costs, funding should be prioritized toward evaluating a program that will be widely relevant to the field or applicable to other sites. Ideally, the program should be scalable and replicable.

- *Organizational capacity and scale.* The program should be well implemented and managed and have the operational capacity to help implement the evaluation. Further, the program should have sufficient participants—or be capable of growing large enough—to support an evaluation.

- *Program stability.* In most cases, programs should demonstrate a track record of stable service delivery, preferably for multiple years. On the other hand, an experimental design can help evaluators assess discrete changes in program design.

- *Adequate funding for program operations and evaluation.* Programs should have sufficient resources to implement the model the research design is testing. This may require additional funding to increase the number of individuals served or to support gathering and entering data.

- *Buy in and research planning participation from staff.* Management, program staff, and administrative staff who will be involved in evaluation implementation must be fully invested in the study and help design implementation and data-collection strategies. Both evaluators and nonprofit staff should see the evaluation as a partnership, have an equal commitment to the study's fidelity and success, and be willing to listen and to contribute to making it work. Experience with implementing performance measurement and prior formative evaluations can be one way to gauge willingness to support additional evaluation work.

The continuum presented here represents an ideal rollout of performance measurement and evaluation activities over the life of a program. Not all programs will follow such a linear, sequential path in building their performance measurement and evaluation activities. Some programs will be able to establish summative evaluation readiness without the foundational work of prior performance measurement and formative evaluation, though that may be harder. Certain programs may leapfrog from performance measurement directly to summative evaluation. Or, a program with no prior history of performance measurement or formative evaluation at all may use summative evaluation as a way to jumpstart these activities. Programs should still meet most of the criteria for evaluability listed above, however, before embarking on a summative evaluation.

Further, performance measurement and evaluation require resources to do well—the availability of resources may limit how fully the continuum, as described here, can be implemented. A high-quality summative evaluation can take a year or more to complete and cost hundreds of thousands of dollars (yet another reason it is not appropriate for every program). Performance measurement and formative evaluations, however, may be more scalable; all but the smallest programs can determine some set of activities that are both feasible and appropriate for assessing and improving performance.

# Conclusion

This brief provides a basic overview of performance measurement and evaluation for nonprofits and funders. As discussed, a program should have a solid theory of change, perhaps accompanied by a logic model, and clearly defined outcomes that will inform and frame performance measurement and evaluation activities. At the earliest stages, the theory of change/logic model may be more theoretical and not yet validated through real-world experience and data. As staff build the capacity to collect performance measurement data and use it to check actual implementation against assumptions and to verify participant outcomes, the nonprofit can further refine the program through a process of continuous improvement. These performance measurement activities can be supplemented by periodic formative evaluations that will allow the nonprofit to make further refinements and gain confidence that the program is functioning as expected. At a later stage, testing the program model against a counterfactual (i.e., what if the program did not exist?) may be appropriate.

These activities represent a performance measurement-evaluation continuum. Nonprofits and funders can use the continuum as a framework for deciding whether, when, and how to apply these methods to assess a program's effectiveness. By understanding what is involved in these different approaches to using data and evidence, nonprofits and funders can better determine what activities are appropriate and most beneficial for programs at various stages of development. By providing a common framework and understanding of the performance measurement-evaluation continuum, this brief can help bridge the divide between nonprofits and funders on how programs should be assessed.

# Glossary of Performance Measurement and Evaluation Terms

Where possible, this performance measurement and evaluation terminology is taken from published resources, but readers should be aware that not all other sources will agree on the specific meanings used here.

*comparison group*—In a quasi-experimental study, a group of people who did not receive program services and who are meant to represent the counterfactual. Since the comparison group is not selected at random, group members may differ from persons who received program services (the treatment group) in both observable and unobservable characteristics which may affect outcomes.

*comparison study/quasi-experimental study*—A type of summative evaluation that measures differences in outcomes for treatment and comparison groups but does not use random selection to assign participants to those groups. Frequently, program participants self-select to either treatment or comparison groups. Most comparison studies will use statistical methods to control for differences in observable characteristics between treatment and comparison subjects, but will generally not be able to control for differences in unobservable characteristics.

*confounder/competing explanation*—Characteristics of evaluation subjects, or other internal or external conditions, that, may affect the outcomes being examined (Jepsen et al. 2004). Confounding characteristics or conditions may offer an alternative reason outcomes changed for program participants. Ideally, an evaluation will attempt to minimize, or control for, any confounders so that changes in outcomes can be more reliably attributed to the program's impact.

*control group*—A cohort of subjects in an evaluation who did not receive the program services being evaluated.

*cost-benefit analysis/cost-effectiveness analysis*—A type of analysis that quantifies how much program services cost, compared to the estimated economic benefits of improved participant outcomes. For example, a cost-benefit analysis of a job training program might compare the cost of its services with savings to the state unemployment insurance fund, based on the program's impact of increasing participants' employment rates.

*counterfactual*—A set of conditions that characterize how participants would have fared had a program not existed. A control group represents the counterfactual in an experimental study (Bangser 2014).

*experimental study/randomized controlled trial*—A type of summative evaluation that selects subjects (e.g., individuals, schools, neighborhoods) and randomly assigns them to one of at least two groups: treatment or control. Such a process helps make the treatment and control groups equivalent—in motivation, ability, knowledge, socioeconomic and demographic characteristics, and so on—at the start of the study and provides an effective counterfactual for measuring program impact (Theodos et al. 2014).

*external validity*—How reliably conclusions about program impact from a study or evaluation can be generalized to different populations, programs, geographies, or time spans (Theodos 2014).

*formative evaluation*—A set of research activities intended to provide information about how a program is being designed or carried out, with the objective of improving implementation and results.[13]

*impact*—The net effect of a program relative to the program never existing. Most typically, impact reflects long-term outcomes, such as persistent changes in participants' situations or behavior. Impacts can be either intended (i.e., those the program meant to cause) or unintended (i.e., those incidental to the program's original goals or objectives).

*input*—A resource applied by a program, such as funds, staff, technology, and materials (Lampkin and Hatry 2003).

*internal validity*—How accurately a study or evaluation measures a program's impact on outcomes for the population it is serving (Theodos et al. 2014).

*logic model*—A graphic that shows how a program is intended to work and achieve results, often based on a specific theory of change. A logic model depicts a path toward a goal, typically including resources, activities, outputs, and outcomes (Knowlton and Phillips 2013).[14]

*natural experiment*—An experimental study in which the selection of treatment and control groups, while not entirely random, is based on some external condition beyond the direct influence of evaluators, program staff, or participants; group composition thus closely resembles random selection. Examples of external conditions can include law or policy changes , differences in laws or policies across jurisdictions, natural disasters, or new technologies.

*observable characteristic*—An attribute or condition that evaluators are able to detect and measure for program participants or evaluation subjects.

*outcome*—An enduring change in participants' attitudes, emotions, knowledge, behavior, health, or social condition brought about by a program's intentional actions. An intermediate outcome is more immediate results (e.g., job training participants who obtain full-time employment). A long-term outcome is a more enduring change further removed from the program's direct results (e.g., participants increase economic well-being) (Lampkin and Hatry 2003, 10).[15]

*output*—The direct product produced by a program or the quantity of work activity completed (Lampkin and Hatry 2003).[16]

*performance management*—A dynamic process designed to explain program operations, monitor outcomes, and ultimately, help programs produce better outcomes for participants. Performance management involves regular, ongoing performance measurement, reporting, analysis, and program modification.[17]

*performance measurement*—Data about a program's operations and outcomes that are collected and analyzed, usually by nonprofit leadership and program staff, to aid in performance management.[18]

*planning study*—A type of formative evaluation that takes place during the design or planning phase to help programs clarify plans and make improvements at an early stage.

*post-then-pre analysis*—An approach to assessing changes in participants, in which evaluators measure an outcome once at the end of the program. Participants are asked to assess retrospectively whether they experienced any change in outcome as a result of the program (Rockwell and Kohn 1989).

*pre-post analysis*—An approach to assessing changes in participants, in which evaluators measure an outcome once at the start of the program and once at the end. Most commonly, the term refers to assessments that do not employ a control or comparison group but examine only changes in outcomes among program participants.[19]

*process study/implementation study*—A type of formative evaluation that looks at the actual implementation of a program. A process study establishes whether a program has attained quantifiable targets and whether it has implemented strategies as specified in the program's logic model (Linnell, Radosevich, and Spack 2002).[20]

*randomization/random selection*—A process of assigning evaluation subjects to treatment and control groups by some mechanism relying on pure chance.

*selection bias*—Differences in evaluation subjects' observable or unobservable characteristics at the point of entry to a program; such characteristics may influence the outcomes being studied (Bangser 2014). While accounting for differences in observable characteristics may be possible through statistical methods, differences in unobservable characteristics, depending on their size and influence, may undermine an evaluation's internal validity. *Self-selection* bias is a specific form of selection bias that occurs when participants are able to choose (i.e., *self-select*) whether they will take part in a program or study. People who choose to participate may be different in some observable or unobservable way from those who choose not to participate, which may influence outcomes being studied.

*self-selection*—A situation whereby individuals are able to choose whether they will take part in a program or study, in contrast to someone assigning them to participate, either randomly or nonrandomly. See also, *selection bias*.

*summative evaluation*—A study that assesses a program's effectiveness in achieving results, based on the program's theory of change or logic model. Depending on the method used, a summative evaluation may be able to determine the program's impact on specific outcomes.

*theory of change*—A conceptual "road map" that outlines how a series of actions can bring about a desired outcome.

*treatment group*—A cohort of subjects in an evaluation who received the program service or set of services being evaluated.

*unobservable characteristic*—An attribute or condition an evaluator is not able to detect or measure for program participants or evaluation subjects.

# Notes

1. "What Is Performance Management?," Perform Well, accessed December 10, 2015, http://www.performwell.org/index.php/performance-management.
2. For more on theories of change, see Hunter (2005) and "What Is Theory of Change?," Center for Theory of Change, 2013, http://www.theoryofchange.org/what-is-theory-of-change/.
3. For more on logic models, see Knowlton and Phillips (2013) and "Why All the Hype about Logic Models?," Nonprofit Answer Guide, 2015, http://nonprofitanswerguide.org/faq/evaluation/why-all-the-hype-about-logic-models/.
4. A related method, *post-then-pre*, relies on a single observation of clients at the end of their participation in the program. Participants are asked to assess retrospectively whether the program has changed their knowledge, attitudes, or behavior in some way. Post-then-pre is meant to address the response-shift bias that can exist in pre-post designs. For more, please see Rockwell and Kohn (1989) and "Quick Tips: Using the Retrospective Post-then-Pre Design," University of Wisconsin Extension, 2005, http://www.uwex.edu/ces/pdande/resources/pdf/Tipsheet27.pdf.
5. For examples of using case management data for performance measurement, see Bogle, Gillespie, and Hayes, 2015.
6. "What's the Difference between Formative and Summative Evaluations?," Austin Independent School District, 2011, https://www.austinisd.org/dre/ask-the-evaluator/whats-difference-between-formative-and-summative-evaluations.
7. Adapted from Susan Allen Nan, "Formative Evaluation."
8. Deborah Linnell, "Process Evaluation vs. Outcome Evaluation," Third Sector New England, 2015, http://tsne.org/process-evaluation-vs-outcome-evaluation. For more on the differences between outcome and process studies, see Linnell, Radosevich, and Spack (2002).
9. For more types of formative evaluation, see Evaluation Toolbox, "Formative Evaluation," 2010, http://evaluationtoolbox.net.au/index.php?option=com_content&view=article&id=24&Itemid=125.
10. For more on deciding when and how to scale up programs, see Bangser (2014).
11. For further discussion of obstacles to conducting a strong experimental study and how to overcome them, see Theodos and coauthors (2014).
12. Adapted from Theodos and coauthors (2014). For additional discussion of evaluability criteria, see Davies (2013), Kaufman-Levy and Poulin (2003), and United Nations Office on Drugs and Crime, "Evaluability Assessment Template," no date, https://www.unodc.org/documents/evaluation/Guidelines/Evaluability_Assessment_Template.pdf.
13. "What's the Difference between Formative and Summative Evaluations?," Austin Independent School District, accessed December 4, 2015, https://www.austinisd.org/dre/ask-the-evaluator/whats-difference-between-formative-and-summative-evaluations.
14. "Why All the Hype about Logic Models?," Nonprofit Answer Guide, accessed December 4, 2015, http://nonprofitanswerguide.org/faq/evaluation/why-all-the-hype-about-logic-models/.
15. "Identify Outcomes," Perform Well, accessed December 4, 2015, http://www.performwell.org/index.php/identify-outcomes.
16. "Terms Used in the Evaluation of Organizational Capacity Development," International Development Research Centre, no date, http://web.idrc.ca/es/ev-43631-201-1-DO_TOPIC.html.
17. "What Is Performance Management?," Perform Well.

18. *Ibid*.

19. Work Group for Community Health and Development, "Chapter 37: Section 4. Selecting an Appropriate Design for the Evaluation," Community Tool Box, accessed December 11, 2015, http://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/experimental-design/main.

20. "What Is the Difference between Process, Outcome and Impact Evaluations?," Nonprofit Answer Guide, 2015, http://nonprofitanswerguide.org/faq/evaluation/difference-between-process-outcome-and-impact-evaluations.

# References

Allen Nan, Susan. 2003. "Formative Evaluation." Boulder, CO: Beyond Intractability, University of Colorado http://www.beyondintractability.org/essay/formative-evaluation.

Bangser, Michael. 2014. "A Funder's Guide to Using Evidence of Program Effectiveness in Scale-Up Decisions." New York: MDRC and Growth Philanthropy Network. http://www.mdrc.org/publication/funder-s-guide-using-evidence-program-effectiveness-scale-decisions.

Bogle, Mary, Sarah Gillespie, and Christopher Hayes. 2015. *Continually Improving Promise Neighborhoods: The Role of Case Management Data.* Washington, DC: Urban Institute.

Davies, Rick.  2013. "Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations." DFID Working Paper 40. Cambridge, UK: Department for International Development.

Dodaro, Gene L. 2011. "Government Performance: GPRA Modernization Act Provides Opportunities to Help Address Fiscal, Performance, and Management Challenges." Testimony before the US Senate Committee on Budget, Washington, DC, March 16. http://www.gao.gov/assets/130/125777.pdf.

Harlem Children's Zone. 2012. "Successful Research Collaborations: Rules of Engagement for Community-Based Organizations." New York: Harlem Children's Zone. http://hcz.org/wp-content/uploads/2014/04/Rules_of_Engagement_paper.pdf.

Hunter, David E. K. 2005. "Daniel and the Rhinoceros." *Evaluation and Program Planning* 29(2): 180–85.

Jepsen, P., S. P. Johnsen, M. W. Gillman, and H. T. Sørensen, 2004. "Interpretation of Observational Studies," *Heart* 90 (8): 956–60. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1768356/.

Kaufman-Levy, Deborah, and Mary Poulin. 2003. *Evaluability Assessment: Examining the Readiness of a Program for Evaluation*. Washington, DC: Juvenile Justice Evaluation Center. http://www.jrsa.org/pubs/juv-justice/evaluability-assessment.pdf.

Knowlton, Lisa Wyatt, and Cynthia C. Phillips. 2013. *The Logic Model Guidebook: Better Strategies for Great Results.* Thousand Oaks, CA: Sage Publications.

Linnell, Deborah, Zora Radosevich, and Jonathan Spack. 2002. *Executive Directors Guide: The Guide for Successful Nonprofit Management*. Boston: Third Sector New England.

Lampkin, Linda M., and Harry P. Hatry. 2003. "Key Steps in Outcome Management." Washington, DC: Urban Institute. http://www.urban.org/research/publication/key-steps-outcome-management.

Linnell, Deborah, Zora Radosevich, and Jonathan Spack. 2002. *Executive Directors Guide: The Guide for Successful Nonprofit Management*. Boston: Third Sector New England.

Mayer, Neil S., Peter A. Tatian, Kenneth Temkin, and Charles A. Calhoun. 2012. "Has Foreclosure Counseling Helped Troubled Homeowners? Evidence from the Evaluation of the National Foreclosure Mitigation Counseling Program." Metropolitan Housing and Communities Center Brief 1. Washington, DC: Urban Institute. http://www.urban.org/research/publication/has-foreclosure-counseling-helped-troubled-homeowners.

Putnam, Kristen. 2004. "Measuring Foundation Performance: Examples from the Field". Oakland: California Healthcare Foundation.

http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/PDF%20M/PDF%20MeasuringFoundationPerformance.pdf.

Rockwell, S. Kay, and Harriet Kohn. 1989. "Post-Then-Pre Evaluation." *Journal of Extension* 27(2). http://www.joe.org/joe/1989summer/a5.php.

Theodos, Brett, Margaret Simms, Rachel Brash, Claudia Sharygin, and Dina Emam. 2014. "Randomized Controlled Trials and Financial Capability: Why, When, and How." Washington, DC: Urban Institute. http://www.urban.org/research/publication/randomized-controlled-trials-and-financial-capability.

## About the Author

**Peter A. Tatian** is a member of the Measure4Change team and a senior fellow in the Metropolitan Housing and Communities Policy Center at the Urban Institute, where he researches housing policy, neighborhood indicators, and community development. He has over 25 years of experience in public policy research and data analysis. He also leads Urban's team providing technical assistance on performance measurement and evaluation to grantees of the US Department of Education's Promise Neighborhoods initiative.

## Acknowledgments

This brief is part of the Measure4Change initiative funded by the World Bank Group. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the author and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine our research findings or the insights and recommendations of our experts. Further information on the Urban Institute's funding principles is available at www.urban.org/support.

The author would like to thank Mary Bogle, Lindsey Buss, Brett Theodos, and Mary Winkler for commenting on earlier versions of this brief.

URBAN INSTITUTE

2100 M Street NW
Washington, DC 20037

www.urban.org

### ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.